Study of Saliency in Objective Video Quality Assessment

Wei Zhang, *Student Member, IEEE*, and Hantao Liu, *Member, IEEE*

***Abstract*— Reliably predicting video quality as perceived by humans remains challenging and is of high practical relevance. A significant research trend is to investigate visual saliency and its implications for video quality assessment. Fundamental problems regarding how to acquire reliable eye-tracking data for the purpose of video quality research and how saliency should be incorporated in objective video quality metrics (VQMs) are largely unsolved. In this paper, we propose a refined methodology for reliably collecting eye-tracking data, which essentially eliminates bias induced by each subject having to view multiple variations of the same scene in a conventional experiment. We performed a large-scale eye-tracking experiment that involved 160 human observers and 160 video stimuli distorted with different distortion types at various degradation levels. The measured saliency was integrated into several best known VQMs in the literature. With the assurance of the reliability of the saliency data, we thoroughly assessed the capabilities of saliency in improving the performance of VQMs, and devised a novel approach for optimal use of saliency in VQMs. We also evaluated to what extent the state-of-the-art computational saliency models can improve VQMs in comparison to the improvement achieved by using "ground truth" eye-tracking data. The eye-tracking database is made publicly available to the research community.**

***Index Terms*— Saliency, video quality assessment, eye-tracking, quality metric, saliency model.**

I. Introduction

THE last few decades have witnessed a phenomenal growth in the use of digital videos in our everyday lives. Video signals, however, are vulnerable to distortion due to causes such as acquisition errors, data compression, noisy transmission channels and the limitations in rendering devices. The ultimate video content received or consumed by the end user largely differs in perceived quality depending on the application. The reduction in video quality may affect viewers' visual experiences or lead to interpretation mistakes in video-based inspection tasks. Finding ways to effectively control and improve video quality has become a focal concern in both academia and industry [1].

Video quality metrics (VQMs), which represent computational models for automatic assessment of perceived video quality aspects, have emerged as an important tool for the optimisation of modern imaging systems [2]. Video quality, to some extent, may be approached as a summation of the quality of individual frames in a video sequence [2], [3]. Therefore, models established for image quality may be reused and extended towards video quality assessment. Taking advantage of sophisticated modelling of image quality and by incorporating the multi-dimensional (i.e., spatial and temporal) structure of video signals, a variety of VQMs have been devised and proven useful in predicting human judgements of video quality [4]–[10]. Yet, notwithstanding the progress made in the development of VQMs, being able to reliably predict the way humans assess the overall video quality or some aspect of it remains an academically rather challenging problem. This is intrinsically due to the fact that our understanding of how video signals and their distortions are perceived by the human visual system (HVS) is still far from complete.

To further enhance the reliability of VQMs, a significant research trend is to investigate the impact of visual attention, which is considered as an essential component of the HVS. Visual attention exists in the HVS as a powerful mechanism that allows effectively selecting the most relevant information from a visual scene [11], [12]. This attentional selection is known to be controlled by two kinds of mechanisms: stimulus-driven, bottom-up mechanism and expectation-driven, top-down mechanism [11]–[13]. In the field of machine vision, visual attention is mainly concerned with the former attentional mechanism, and is often interchangeably referred to as *saliency* [14]–[16]. The empirical foundation of saliency modelling lies in the eye movements of human observers, intent on explicitly addressing fixations during free-viewing of a visual stimulus [14], [17]. A computational model of saliency generally outputs a topographic map that represents conspicuousness of scene locations [18]. To incorporate saliency aspects in VQMs, the vast majority of existing approaches have focused on simply using a specific saliency model to weight the local distortions measured by a specific VQM [19]–[27]. For example, in [21], a well-established saliency model (i.e., SaliencyToolBox [28]) is integrated into two popular VQMs (i.e., SSIM and MSE [29]) to improve their performance for the assessment of packet-loss-impaired video. In such an approach, the evaluation of the benefits of saliency (e.g., as the results reported in the studies in [19]–[27]) may heavily depend on the reliability of the saliency model used. Fundamental problems such as how saliency plays a role in judging video quality and how to integrate saliency into VQMs in a perceptually optimised way

Manuscript received April 12, 2016; revised September 19, 2016 and November 23, 2016; accepted December 26, 2016. Date of publication January 9, 2017; date of current version January 30, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rahul Vanam.

The authors are with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K. (e-mail: zhangw71@cardiff.ac.uk; hantao.liu@cs.cardiff.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2651410

remain unsolved. To investigate these topics, eye-tracking data that represent "ground truth" saliency in the particular context of video quality are highly desirable.

II. Related Work and Contributions

A. Related Work

Eye-tracking studies have been attempted to understand saliency in relation to video quality assessment [30]–[33]. The study in [30] focuses on investigating the relative impact of artifacts in the region of interest (ROI) and that in the background region on the overall video quality. ROI was determined by means of eye-tracking experiments. It shows that the quality of the ROI is about ten times more important for the overall quality judgement than the quality of the background. A subjective experiment was conducted in [31] to exam the impact of visual saliency on the annoyance of packet loss distortion. Saliency was identified using free-viewing eye-tracking data. The results show that distortions in salient regions are perceived significantly more annoying than that in the non-salient regions. In [32], eye-tracking experiments were performed with the aim to understand whether a quality scoring task can affect the deployment of fixations. The study indicates that the scoring task given to the subjects may have an impact on where they look in videos. It also demonstrates that adding eye-tracking data collected under a quality scoring task into VQMs does not significantly improve their performance. The findings sufficiently support the high relevance of saliency to VQMs and the importance of collecting eye-tracking data under free-viewing conditions.

In general, psychophysical studies as mentioned above strongly imply that visual saliency plays a vital role in judging video quality. Due to the "ground truth" nature of eye-tracking, modelling saliency in VQMs largely relies on the availability of a dedicated and reliable eye-tracking database. However, existing eye-tracking data relevant to video quality are limited with respect to the number of human subjects, the number of stimuli and the degree of stimulus variability. For example, the eye-tracking experiments reported in [30]–[32] all made use of a single type of distortion (i.e., H.264 compression artifacts), which affects the validity of the results in terms of generalisation. The other drawback to existing eye-tracking data is that they are potentially biased due to the methodology used for data collection. More specifically, in their eye-tracking experiments each observer had to view the same scene repeated several times (with multiple types and/or levels of distortion). In such a scenario, the viewers might be forced to e.g. learn to look for the artifacts rather than observing the stimuli naturally. As a consequence, the recorded fixations might be more affected by the visual distortions rather than the natural scenes. Such involvement of stimulus repetition (i.e., repeated versions of the same scene) and its implications for observers' perception challenge the reliability of existing eye-tracking data. This kind of bias is increasingly noticed as a general challenge to subjective testing, where subjects interact with the same stimuli repeatedly. Effort has been made to refine traditional experimental methods, such as for scoring video or speech quality [34]. It is worth investigating a refined method for eye-tracking.

Due to the absence of sufficient eye-tracking data, studies integrating saliency into VQMs in a perceptually meaningful way are still very limited. A fundamental question remains whether it is natural scene saliency (i.e., saliency derived from the original, non-degraded content of a natural scene, and referred to as NSS) or distorted scene saliency (i.e., saliency derived from a visual scene distorted with artifacts, and referred to as DSS) that should be included in VQMs. Due to the lack of sound evidence to guide choice, researchers often make an *ad hoc* decision by either generating saliency from the reference videos (e.g., [20], [23], [35]) or from the distorted videos (e.g., [21], [24], [32]). Such a rather random selection of saliency (i.e., NSS or DSS) runs the risk of compromising the effectiveness of the inclusion of saliency in VQMs. Determining optimal use of saliency in VQMs is worth further investigation.

B. Contributions of the Paper

1) Eye-tracking data for video quality research are already available in the literature. However, they are either strongly biased or limited by their scale to be able to produce statistically sound findings. We aim to build a large-scale and reliable eye-tracking database. To this end, we focus on refining traditional experimental methodologies and developing an alternative methodology for reliably recording fixations of videos of varying quality. The refined methodology is rigorously validated and can be used as a generic framework for studying saliency in video quality assessment. Moreover, we have made the eye-tracking database publicly available [36] to facilitate research on modelling saliency in VQMs.
2) On the basis of the "ground truth" eye-tracking data, dedicated analysis is performed to better understand human fixation behaviour. New findings are achieved regarding the differences in fixation deployment when viewing the original versus distorted scenes and when viewing the static versus dynamic scene.
3) So far, there is no reliable, scientifically sound evidence on whether it is NSS or DSS should be included in VQMs. With both NSS and DSS reliably measured in our eye-tracking experiments, we aim to clarify the knowledge on the intrinsic added value of both types of saliency in VQMs. We found that the benefit of adding NSS to VQMs was marginal, but DSS could improve the VQMs' performance to a considerable extent.
4) To build a benchmark for saliency-based VQMs, the "ground truth" DSS is then added to several best-known VQMs in the literature. We aim to provide accurate quantitative evidence, by means of an exhaustive statistical evaluation, on to what extent saliency can actually benefit VQMs depending on the distortion types assessed and the VQMs used.
5) On the basis of DSS, we further investigate combining local distortions and their corresponding saliency. Rather than focusing on a VQM-specific integration approach, we devise a generic approach for perceptually optimising the use of saliency in VQMs.

6) We also evaluate thoroughly to what extent state-of-the-art saliency models can improve the performance of VQMs compared to improvement achieved by using eye-tracking data. Many saliency models are available in the literature (see e.g., in [12]); but the general applicability of these models in VQMs is not fully justified. The results of the quantitative comparison serve as a reference for pre-screening saliency models for the particular application domain of video quality.

III. Proposed Experimental Methodology

A. Refined Experimental Design

Unlike previous studies which are potentially biased due to the subjects experiencing massive stimulus repetition, our proposed methodology includes dedicated control mechanisms to eliminate such bias. In addition, our experiment contains a large degree of stimulus variability in terms of video content, distortion type as well as degradation level. This yields a large-scale database, involving 160 human observers, 160 video sequences, and 3200 eye-tracking trials.

1) Stimuli: The test stimuli were taken from the LIVE video quality database [37]. The database is formed of 10 uncompressed, high-quality source/reference videos with a wide variety of content, and a set of 150 distorted videos (i.e.,15 distorted videos per reference) of four different distortion types, namely MPEG-2 compression (i.e., referred to as MPEG-2), H.264 compression (i.e., referred to as H.264), simulated transmission of H.264 compressed bit streams through error-prone IP networks (i.e., referred to as IP) and through error-prone wireless networks (i.e., referred to as Wireless). Per video, a difference mean opinion score (i.e., DMOS) was generated from an extensive subjective quality assessment study.

2) Protocol: A quality assessment database typically involves deliberate stimulus repetition, where a reference video exists simultaneously with a number of its distorted versions of varying quality. In the literature, eye-tracking experiments are commonly conducted using a "within-subjects" design, in which the same group of subjects views all stimuli [31]–[33]. This methodology, however, potentially contaminates the results due to carryover effects, which refers to any effect that carries over from one experimental condition (i.e., viewing a stimulus) to another (i.e., viewing another stimulus originated from the same reference) [38], [39]. In our experiment, each reference video corresponds to 16 variations (i.e., 15 distorted + 1 original), which makes data collection prone to undesirable effects such as fatigue, boredom and learning from practice and experience, and consequently increases the chances of skewing the experimental results. To improve the reliability of data collection, we propose to adopt an alternative methodology, namely "between-subjects" [40], in which multiple groups of subjects are randomly assigned to partitions of stimuli, each contains little or no stimulus repetition.

3) Experimental Procedure: The test dataset was divided into 8 partitions of 20 videos each, and only two repeated versions of the same scene were allowed in each partition. To further reduce the carryover effects, each session per subject was divided into two sub-sessions with a "washout" period in between; and by doing so, each subject viewed 10 videos (i.e., half partition) without stimulus repetition in a separate session. Additional mechanisms were applied to control the order in which participants per group perform their tasks: (1) half of the participants viewed the first half of the stimuli first, and half of the participants viewed the second half first; (2) the stimuli in each sub-session were presented to each subject in a random order. A dedicated control was also added to deliberately include a mixture of all distortion types and the full range of distortion levels in each sub-session.

A standard office environment as specified in [41] was set up for the conduct of our experiment. The stimuli were displayed on a 19-inch LCD monitor with a native resolution of 1024×768 pixels. The viewing distance was approximately 60cm. Eye movements were recorded using an image-processing-based contact-free tracking system with sufficient head movement compensation (SensoMotoric Instrument (SMI) Red-m). The eye-tracking system featured a sampling rate of 120Hz, a spatial resolution of 0.1 degree and a gaze position accuracy of 0.5 degree. Before the start of the actual experiment, each participant was provided with instructions on the procedure (e.g., the task, the format of stimuli and the timing) of the experiment. A training session was conducted as a full-scale rehearsal in order to familiarise the participant with the experiment. The video stimuli used in the training session were different from those used in the real experiment. Each full session per subject consisted of two successive sub-sessions with a break of 60 minutes between sub-sessions. Each individual sub-session was preceded by a 9-points calibration of the eye-tracking equipment. The participants were instructed to experience the videos in a natural way ("view it as you normally would"). Each video was displayed followed by a mid-gray screen lasting 3 seconds.

We recruited 160 participants from university students and staff members, including 80 males and 80 females with their ages ranging from 19 to 42. They were all inexperienced with video quality assessment and eye-tracking. The subjects were not tested for vision defects, and we considered their verbal expression of the soundness of their own vision was sufficient. The participants were first randomly divided into 8 groups of equal size, each with 10 males and 10 females; and the 8 groups of subjects were then randomly assigned to 8 partitions of stimuli. This gives a sample size of 20 subjects per test stimulus.

4) Saliency Map: Saliency that represents stimulus-driven, bottom-up visual attention is derived from free-viewing fixations [42], [43]. Fixations were extracted using the *SMI BeGaze Software* with minimum fixation duration threshold set to 100ms. A fixation was defined by SMI's Software using the dispersal and duration based algorithm established in [44]. For a given video sequence, a topographic saliency map per frame is constructed by accumulating fixations over all subjects (i.e., 20 in our experiment) and with each fixation location giving rise to a gray-scale patch that simulates the foveal vision of the HVS [30], [32], [33]. The activity of the patch is modelled as a Gaussian distribution, of which the width σ approximates

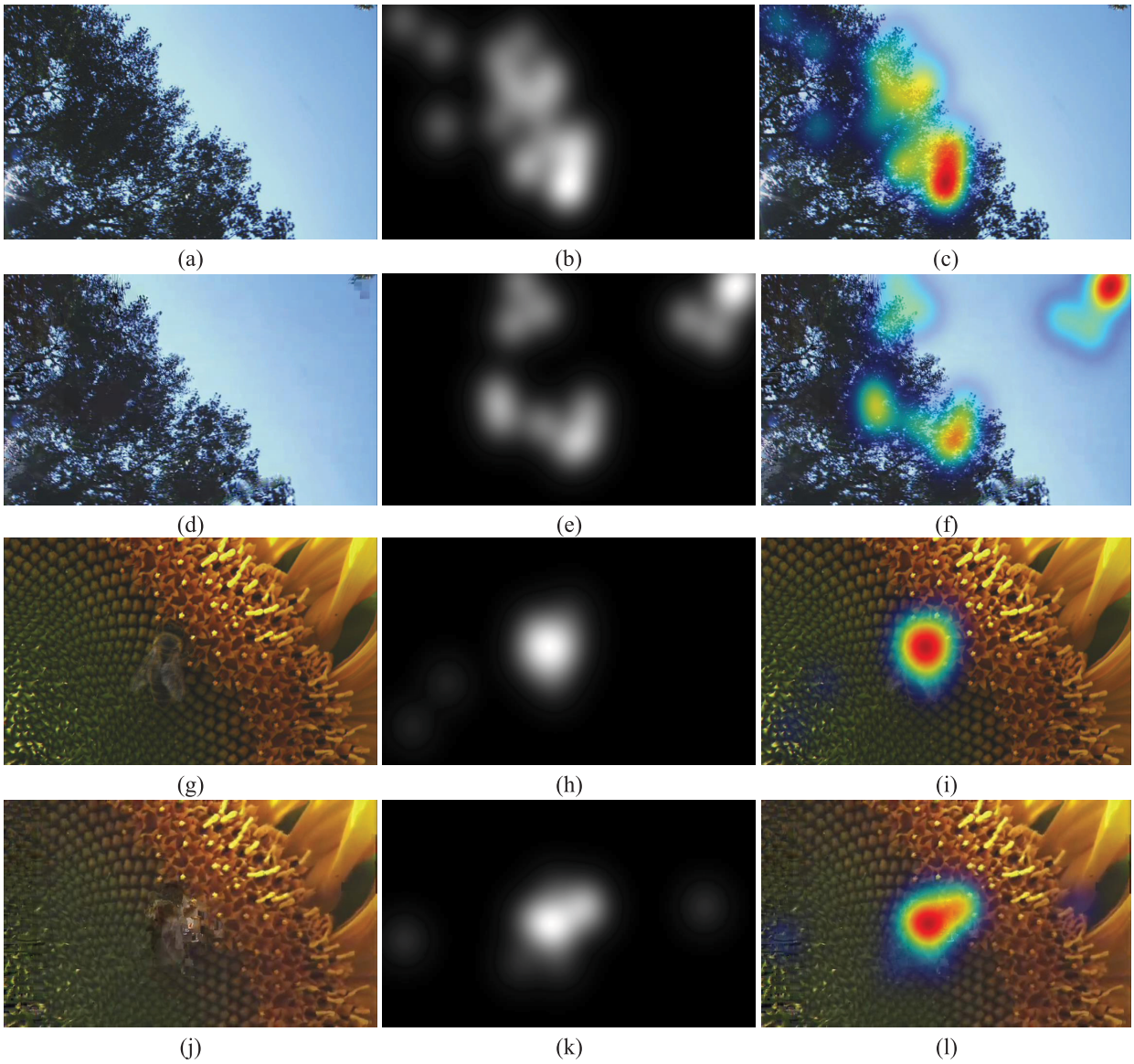


Fig. 1. Illustration of saliency map for a frame taken from an original video and saliency map for the distorted version of the same frame, for two different sample scenes in our experiment. (a) and (g) are original frames. (d) and (j) are distorted frames. (b), (e), (h) and (k) are saliency maps of (a), (d), (g) and (j). (c), (f), (i) and (l) are corresponding heatmaps.

the size of the fovea (i.e., 2° of visual angle, 45 pixels width in our experiment). The saliency map (SM) is calculated as:

$$SM(x, y) = \sum_{i=1}^{N} \exp[-\frac{(x_i - x)^2 + (y_i - y)^2}{\sigma^2}] \quad (1)$$

where (x_i, y_i) indicates the spatial coordinates of the ith fixation, N is the total number of fixations. The intensity of the resulting saliency map is linearly normalised to the range [0, 1]. We follow conventional practice of relevant studies [30]–[33]: when there is no experimental error (e.g., participants failing to complete the entire trial or interrupted data recording due to system failure), all recorded eye-tracking data are deemed valid. Outlier detection may be applied to the dataset. It should be noted that determining whether or not an observation (e.g., fixation) is an outlier is ultimately a subjective exercise [45], and rejection of outliers may be acceptable e.g., when the distribution of measurement error is confidently known [46]. Considering there is no rigid definition of what constitutes an outlier [45], we decide to retain all recorded fixations for further analysis. Fig. 1 illustrates two different sample scenes; and for each scene it shows first the measured saliency map for a representative frame taken from the reference video and then for the corresponding frame from the distorted video (note that saliency maps for the entire database can be accessed via [36]).

B. Validation: Proposed Reliability Testing

Eye-tracking data recorded for the purpose of visual quality research strongly differ in their reliability depending on the choices made in the experimental methodology, such as the sample size and the way of presenting stimuli to observers [47], [48]. Therefore, to be able to draw upon

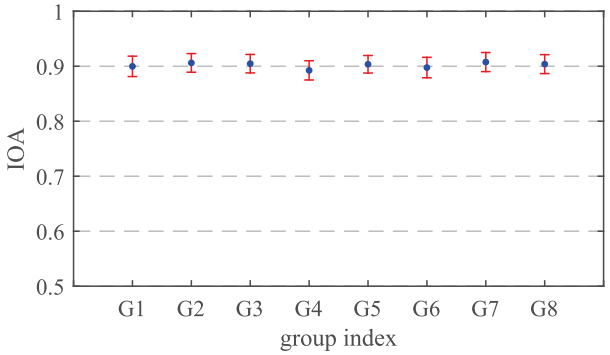


Fig. 2. Illustration of the inter-observer agreement (IOA) value averaged over all stimuli assigned to each subject group in our experiment. The error bars indicate the 95% confidence interval.

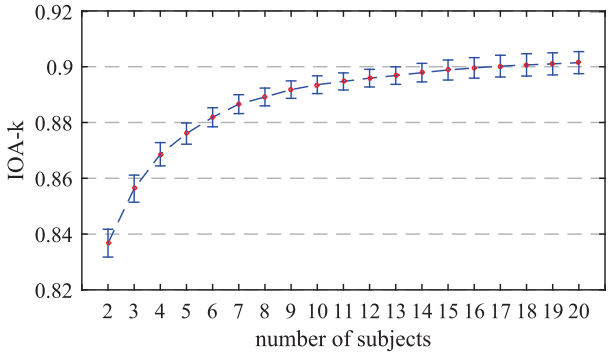


Fig. 3. Illustration of the inter-k-observer agreement (IOA-k) value averaged over all stimuli contained in our dataset. The error bars indicate the 95% confidence interval.

eye-tracking data as a solid "ground-truth," it is crucial to rigorously validate the reliability of the collected data. We propose and perform systematic reliability testing to assess (1) whether the variances in the eye-tracking data obtained from different subject groups are consistent; and (2) whether the sample size is adequate.

1) Homogeneity of Variances Between Groups: Since a between-subjects methodology is employed, it is important to know whether the variances of eye-tracking data across all subject groups are homogeneous. To be able to identify such homogeneity, we measure the inter-observer agreement (IOA), which refers to the degree of agreement in saliency among observers viewing the same stimulus [49], [50]. In our implementation, IOA is quantified per frame by comparing the saliency map generated from the fixations over all-except-one observers to the saliency map built upon the fixations of the excluded observer; and by repeating this operation so that each observer serves as the excluded subject once. The similarity between two saliency maps is measured by the widely used area under the receiver operating characteristic curve (AUC) [12]. Note that alternative similarity measures to compare saliency maps do exist (e.g., Pearson linear correlation coefficient (CC) and normalized scanpath saliency (NSpS)), but since conclusions tend to be consistent over these measures [12], [51], we decided to focus on AUC only. The per-frame IOA is averaged over all frames of a video to generate the per-video IOA: the larger the IOA value, the smaller the variation in fixations among viewers, thus the more reliable the eye-tracking data. Fig. 2 illustrates the per-video IOA averaged over all video stimuli assigned to each subject group in our experiment. It shows that the IOA remains very similar across eight subject groups. A statistical significance test (i.e., analysis of variance (ANOVA)) is performed and the results (i.e., $P > 0.05$ at 95% confidence level) show that there is no statistically significant difference between groups. The above evaluation indicates that a high degree of consistency across groups is found in our eye-tracking data.

2) Data (Saliency) Saturation: To determine the sample size for an eye-tracking experiment, researchers either follow the rule of thumb (i.e., use of 5-15 participants [48]) or use "data saturation" as a guiding principle to make sure a given/chosen sample size is sufficient to cause a "saturated" saliency map. The latter means a saliency map reaches the point at which no new information is observed. We evaluate whether the sample size is adequate to reach such "saturated" saliency (i.e., a proxy of sufficient degree of reliability) in our data. The validation is again based on the principle of IOA, which is extended to an inter-k-observer agreement measure (i.e., referred to as IOA-k, and $k = 2, 3, \ldots, 20$ in our case). More specifically, for a given stimulus, IOA-k is calculated by randomly selecting k observers among all. Fig. 3 illustrates the IOA-k value averaged over all video stimuli in our entire dataset. It shows that "saturation" occurs with 18 participants, although a reasonably high degree of consistency in fixation patterns is already reached with 15 participants. It demonstrates that our chosen number of 20 observers (per subject group and thus per stimulus) is fairly sufficient to yield stable/saturated saliency maps.

C. Behaviour Analysis: Fixation Deployment

1) Original Versus Distorted Scenes: Fig. 1 visualises typical correspondences and differences in saliency between the reference and its distorted scene (i.e., NSS and DSS). In general, there exist consistent patterns between NSS and DSS maps, e.g., the highly salient regions tend to occur around the same places. However, there are some observed deviations, which are seemingly caused by the appearance of distortion. The visible artifacts occurring at the top-right corner in Fig. 1(d) seems to cause an obvious change in saliency (e.g., see the difference between Fig. 1(b) and (e)). This may be due to the distraction power of the localised artifacts is so strong that it offsets the deployment of NSS. In Fig. 1(j), some annoying artifacts happen to occur around the salient object (i.e., the bee in the centre) in the scene, which only leads to a slight deviation in saliency relative to its original pattern (e.g., see the difference between Fig. 1(h) and (k)). It is worthwhile to better understand how saliency deployment is affected by the presence of visual distortion. Such knowledge would provide a grounding for the perceptual integration of saliency and VQMs. We further investigate the observed tendencies in the changes of saliency induced by distortion. More specifically, we evaluate the impact of both distortion strength and distortion type on the deployment of saliency.

TABLE I

NSS-DSS SIMILARITY (MEASURED BY AUC, NSpS AND CC) FOR DIFFERENT LEVELS OF VISUAL QUALITY: EXCELLENT, GOOD, FAIR AND POOR

	MEAN±STD			
	Excellent	Good	Fair	Poor
AUC	0.815±0.058	0.810±0.059	0.809±0.053	**0.781±0.030**
NSpS	1.928±0.921	1.942±0.920	1.881±0.843	**1.367±0.425**
CC	0.657±0.138	0.647±0.134	0.649±0.126	**0.560±0.082**

TABLE II

NSS-DSS SIMILARITY (MEASURED BY AUC) FOR DIFFERENT DISTORTION TYPES AND FOR DIFFERENT VIDEO SCENES

Scene	MEAN±STD			
	IP	Wireless	H.264	MPEG-2
pa	0.815±0.076	0.826±0.059	0.822±0.063	0.819±0.056
rb	0.748±0.084	0.764±0.085	0.754±0.083	0.758±0.094
rh	0.765±0.082	0.752±0.079	0.763±0.080	0.775±0.074
tr	0.811±0.060	0.803±0.057	0.810±0.057	0.806±0.059
st	0.874±0.051	0.879±0.058	0.878±0.048	0.881±0.051
sf	0.920±0.040	0.920±0.039	0.925±0.038	0.922±0.037
bs	0.769±0.082	0.762±0.079	0.791±0.064	0.773±0.070
sh	0.749±0.092	0.761±0.084	0.747±0.088	0.763±0.081
mc	0.758±0.084	0.761±0.077	0.764±0.080	0.759±0.075
pr	0.830±0.077	0.843±0.060	0.833±0.073	0.836±0.073
ALL	**0.786±0.097**	0.806±0.088	0.805±0.089	0.806±0.087

For each distorted frame in the dataset, we quantify the difference between a DSS map and its corresponding NSS map using three popular similarity measures: AUC, NSpS and CC as mentioned in Section III-B. The use of these measures is already described in more detail in [52], and their general meaning in our context is as follows: when AUC$>$0.5 or NSpS$>$0, the higher the value of the measure the more similar NSS and DSS are; when CC is close to -1 or 1, the similarity between NSS and DSS is high, when CC is close to 0, the similarity is low. Our evaluation is based on all data points (i.e., all individual frames of 150 distorted video stimuli) of NSS-DSS similarity calculated by AUC, NSpS and CC. To investigate the effect of distortion strength on NSS-DSS similarity, video stimuli are categorised into four levels of visual quality by dividing the full range of DMOS into four equal intervals. This reflects four levels of quality: "Excellent", "Good," "Fair" and "Poor" as also studied in [37]. Table I illustrates the NSS-DSS similarity averaged for four quality levels. It tends to show that the degree of NSS-DSS similarity decreases as the distortion strength increases. It reveals a significant drop in NSS-DSS similarity at low visual quality, which implies that the distraction power of strong distortions may come into significantly impact the perception of the natural scene.

The impact of distortion type on NSS-DSS similarity (in terms of AUC; NSpS and CC exhibit the same trend of changes and thus are not included in the table) is illustrated in Table II, where video stimuli are categorised according to the source of distortion (i.e., Wireless, IP, H.264 and MPEG-2) in the LIVE database. We also further breakdown the grouping into the per-scene level, resulting in four average AUC values for each visual scene. It tends to show that IP distortion produces a larger extent of saliency deviation between NSS and DSS than other three distortion types. On average, compared to Wireless, H.264 and MPEG-2, IP distortion yields a smaller mean AUC with a larger standard deviation. This is probably due to the difference in the perceptual characteristics between distortion types. IP distortion usually appears as a surprising "glitch" in a fairly large area in a scene [37], which may cause considerable distraction during viewing the scene; whereas Wireless, H.264 and MPEG-2 often generate less surprising distortion, such as the uniformly distributed artifacts throughout the entire scene or some localised artifacts in small regions [37]. Table II also shows that NSS-DSS similarity seems to be affected by scene content, e.g., a large AUC with small standard deviation is consistently found for the scene "sf" (i.e., the scene shown in Fig. 1(g) and (j)). This may be explained by the fact that the scene contains a highly salient object that dominates the distribution of fixations, and that the contribution of distortion to the deployment of saliency is relatively small.

2) Static Versus Dynamic Scenes: It should be noted in Fig. 1 that the saliency map does not represent the saliency of a certain frame when being viewed as an independent static picture. The fixations per frame were actually collected when observers viewing the context of motion picture, and therefore, the per-frame saliency map contains both spatial and temporal aspects of visual perception. We further explore human behavioural responses to static and dynamic scenes; and investigate saliency deployment under both contexts. To this end, we conducted an eye-tracking experiment, where 20 subjects were recruited to view freely some sample frames taken from our video stimuli. We limited the study to the undistorted stimuli only in order to avoid introducing an additional variable (i.e., distortion) to the experiment. Two representative frames were extracted from each reference video, resulting in a total of 20 static stimuli. Each participant viewed each stimulus for 10 seconds with the same experimental setup as described in Section III.

Fig. 4 illustrates the comparison of saliency collected for the same frame when viewed as part of a video sequence and as an independent static scene. It clearly shows the deviations in saliency deployment: under the situation of viewing a static scene, fixations tend to cluster around salient objects, such as text and faces; however, when the same scene is placed in the video context, fixations are more affected by dynamics of the sequence, e.g. the movement of an object. To quantify such difference, we calculate the AUC between the two kinds of saliency for each stimulus pair. To make a rigorous comparison, we also vary the duration used for generating a saliency map for the "static" case, covering the intervals of 0-50ms, 0-200ms, 0-500ms till 0-10s. Fig. 5 illustrates the similarity in saliency between the "static" and "dynamic" conditions. In general, it shows a noticeable difference (i.e., AUC is around 0.8) independent of the viewing time used for the "static" case.

IV. THE INTEGRATION OF NSS VERSUS DSS IN VQMS: A COMPARATIVE ANALYSIS

As described in Section II-A, it is still unclear whether it is NSS or DSS that should be included in the design of

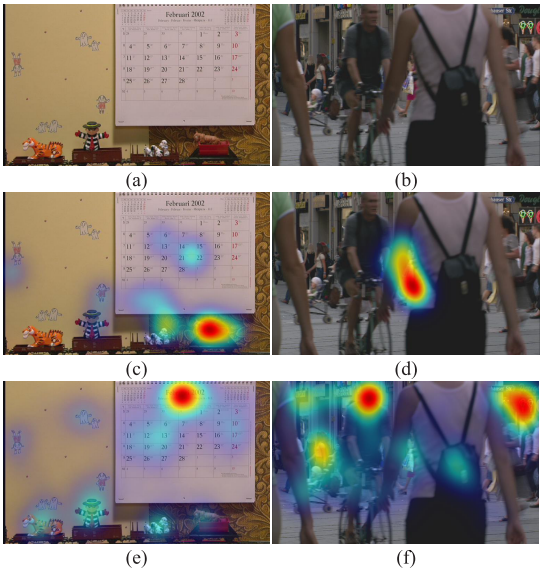


Fig. 4. Illustration of the comparison of saliency for the same scene (a) or (b) when being viewed as part of a video sequence (c) or (d) and as an static picturee (e) or (f).

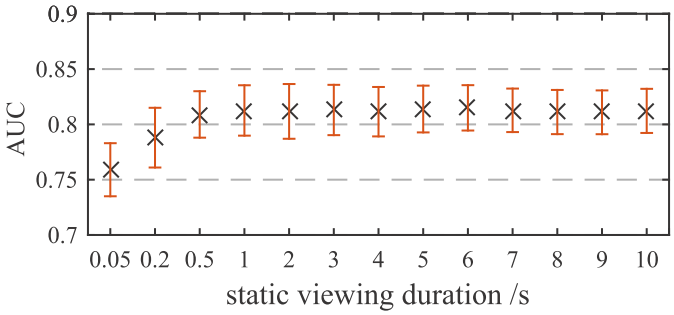


Fig. 5. Illustration of the fixation deployment similarity (measured by AUC) between "static" and "dynamic" viewing conditions. Errorbars indicates 95% confidence level.

saliency-based VQMs. It is important to understand whether the observed difference between NSS and DSS (as detailed in Section III-C) is sufficiently large to actually affect the performance gain for VQMs. To this end, we simply add both types of saliency to several well-established VQMs in the literature, and compare the performance gain obtained when including NSS versus DSS.

A. Evaluation Framework

1) VQMs: Eight widely recognised full-reference VQMs, namely PSNR, SSIM, ViS3, STMAD, spatial MOVIE, temporal MOVIE, MOVIE and VSSIM are applied in our evaluation:

PSNR: The peak signal-to-noise ratio is based on the mean squared error between the distorted video and its reference on a pixel-by-pixel basis.

SSIM: The structural similarity index [29] measures per-frame quality of a video based on the degradation in structural information. The SSIM is first calculated frame-by-frame on the luminance component of the video and then averaged over all frames to achieve an overall quality prediction.

ViS3: The ViS3 algorithm [9] contains two stages: the first stage measures the quality based on spatial distortion, and the second stage measures the quality based on the dissimilarity between spatiotemporal slice images. The overall video quality prediction is a combination of the quality scores calculated at two stages.

STMAD: The spatiotemporal MAD [8] extends the image-based quality metric MAD by taking into account the visual perception of motion artifacts. The motion artifacts are measured on the time-based slices of the original and distorted videos. The velocity of moving objects is taken into consideration to adjust the locally measured degradations.

MOVIE: The motion-based video integrity evaluation index [7] utilises a general, spatio-spectrally localized multi-scale framework for evaluating dynamic video fidelity. It integrates both spatial and temporal (and spatio-temporal) aspects of distortion assessment, resulting in three different versions of the MOVIE index, namely the spatial MOVIE (SMOVIE), the temporal MOVIE (TMOVIE) and MOVIE.

VSSIM: The video SSIM [10] is an improved version of the single-scale SSIM taking into account the motion perception of the HVS.

The VQMs above range from the purely pixel-based VQMs without characteristics of the HVS (i.e., PSNR) to VQMs that contain rather complex HVS properties (i.e., VQMs under test except for PSNR). Some VQMs operate on a frame-by-frame, spatial-distortion-only basis (i.e., PSNR, SSIM, SMOVIE), whereas other VQMs predict local, spatio-temporal quality by taking into account several frames of the sequence (i.e., ViS3, STMAD, TMOIVE, MOVIE, VSSIM). All VQMs result in a quantitative per-frame distortion map (PFDM) which represents a local quality degradation profile. Note that other well-known VQMs that do not explicitly produce a PFDM, such as VQM software tool [6], are not included in our study. Integrating a (per-frame) saliency map into such kind of VQMs is not straightforward and is, therefore, outside the scope of this paper. Also, reduced-reference and no-reference VQMs are not included, since they are still in the early stages of development and remain limited in their sophistication, which makes studying the added value of saliency in these VQMs less meaningful.

2) Saliency-Based VQMs: Saliency map (SM), either NSS or DSS, is integrated into a VQM via locally weighting (i.e., by multiplying) the PFDM with the corresponding SM per frame (of size M×N pixels), yielding a saliency weighted PFDM (SW-PFDM):

$$SW - PFDM = \frac{PFDM(x,y) * SM(x,y)}{\sum_{x=1}^{M}\sum_{y=1}^{N} SM(x,y)} \tag{2}$$

where PFDM is measured by an VQM, SM is generated from our eye-tracking data. Once the PFDM is upgraded to the SW-PFDM, the remaining operations of the VQM proceed as usual to produce an overall quality score. It should be noted that PFDM and SM are simply combined in our implementation. This simple weighting has been conventionally used in the literature [19]–[21], due to its nature of being parameter free and universally applicable. A more sophisticated

TABLE III

COMPARISON OF PERFORMANCE (CC) FOR DIFFERENT VQMS AND THEIR CORRESPONDING SALIENCY-BASED VERSIONS. VALUES IN THE BRACKETS REPRESENT THE PERFORMANCE GAIN (I.E., THE INCREASE IN CC (ΔCC)) OF A SALIENCY-BASED VQM OVER ITS ORIGINAL VERSION

	PSNR	SSIM	ViS3	STMAD	SMOVIE	TMOVIE	MOVIE	VSSIM
Original	0.539	0.500	0.826	0.823	0.740	0.823	0.795	0.584
DSS-based	0.567(0.028)	0.544(0.044)	0.839(0.013)	0.834(0.011)	0.771(0.031)	0.836(0.013)	0.829(0.034)	0.605(0.021)
NSS-based	0.562(0.023)	0.504(0.004)	0.829(0.003)	0.827(0.004)	0.734(-0.006)	0.821(-0.002)	0.790(-0.005)	0.581(-0.003)
RSS-based	0.542(0.003)	0.503(0.003)	0.817(-0.009)	0.816(-0.007)	0.731(-0.009)	0.815(-0.008)	0.788(-0.007)	0.580(-0.004)

TABLE IV

COMPARISON OF PERFORMANCE (SROCC) FOR DIFFERENT VQMS AND THEIR CORRESPONDING SALIENCY-BASED VERSIONS. VALUES IN THE BRACKETS REPRESENT THE PERFORMANCE GAIN (I.E., THE INCREASE IN SROCC (ΔSROCC)) OF A SALIENCY-BASED VQM OVER ITS ORIGINAL VERSION

	PSNR	SSIM	ViS3	STMAD	SMOVIE	TMOVIE	MOVIE	VSSIM
Original	0.523	0.525	0.816	0.825	0.727	0.806	0.789	0.587
DSS-based	0.543(0.020)	0.567(0.042)	0.833(0.017)	0.830(0.005)	0.755(0.028)	0.808(0.002)	0.809(0.020)	0.606(0.019)
NSS-based	0.536(0.013)	0.520(-0.005)	0.837(0.021)	0.826(0.001)	0.685(-0.042)	0.801(-0.005)	0.776(-0.013)	0.574(-0.013)
RSS-based	0.509(-0.014)	0.536(0.011)	0.811(-0.005)	0.818(-0.007)	0.724(-0.003)	0.799(-0.007)	0.792(0.003)	0.571(-0.016)

combination strategy may further improve the performance of a specific VQM (see, e.g., [35]), but is often incompatible with other VQMs. Using such VQM-specific weighting would make the comparative study of VQMs impractical and less meaningful. Since there does not exit a more sophisticated and generic weighting strategy, we decide to use the simple weighting to ensure a fair comparison between NSS and DSS.

3) Performance Evaluation Criteria: As prescribed by the Video Quality Experts Group [53], the performance of a VQM is quantified by the Pearson linear correlation coefficient (CC) and the Spearman rank order correlation coefficient (SROCC) between the quality predictions of the VQM and the DMOS scores. Seemingly, the quality assessment community is accustomed to fitting the predictions of a VQM to the DMOS scores [53]. A nonlinear mapping may, e.g., account for a possible saturation effect in the quality scores at high quality. It usually yields higher correlations in absolute terms, while generally keeping the relative differences between VQMs [54]. As also explained in [55], without a sophisticated nonlinear fitting the correlations cannot mask a bad performance of the VQM itself. To better visualise differences in performance, we avoid any nonlinear fitting and directly calculate correlations between the VQM's predictions and the DMOS scores.

B. Comparison of NSS Versus DSS Applied in VQMs

Based on the "ground truth" NSS and DSS obtained from our eye-tracking experiments, we set out to evaluate to what extent adding both types of saliency is beneficial for the prediction performance of VQMs. We compare the performance gain that can be obtained when adding NSS versus DSS to PSNR, SSIM, ViS3, STMAD, SMOVIE, TMOVIE, MOVIE or VSSIM. To better visualise the comparison, we also include the so-called random scene saliency (RSS), which provides a baseline for the performance gain when adding saliency to VQMs. RSS is generated for a video stimulus by randomly selecting saliency from our collection of NSS and DSS.

Table III and IV summarise the performance of VQMs in terms of CC and SROCC. Each entry represents the performance of a VQM (with or without saliency weighting) on the entire LIVE video quality database (i.e., a total of 150 data points/distorted video stimuli). In general, both tables demonstrate that the performance of all VQMs is consistently enhanced by including DSS. The gain in their performance ranges from 0.011 to 0.044 in CC and 0.002 to 0.042 in SROCC. On the contrary, adding NSS or RSS does not seem to be beneficial for VQMs. The performance gain is either marginal (non-existent) or even negative, e.g., adding NSS and RSS to SSIM corresponds to an increase of 0.004 and 0.003 in CC respectively; both NSS and RSS deteriorate the performance of MOVIE (i.e., NSS causes a decrease of 0.005 in CC, and RSS decreases CC by 0.007). Fig. 6 plots the overall performance gain (i.e., expressed by the increase in CC (ΔCC)) that can be obtained by adding three types of saliency to all VQMs. On average, incorporating DSS yields a promising gain for VQMs (i.e., $< \Delta CC >= 0.024$); and VQMs do not actually profit from being extended with NSS (i.e., $< \Delta CC >= 0.002$) or RSS (i.e., $< \Delta CC >= -0.005$). To check such effects with a statistical analysis, a nonparametric test (i.e., Wilcoxon signed rank test [56]) analogue to a paired samples t-test (as ΔCC values are tested to be not normally distributed) is performed once between DSS and NSS and once between DSS and RSS. The test results show that DSS weighted VQMs receive statistically significantly higher performance gain than NSS or RSS weighted VQMs

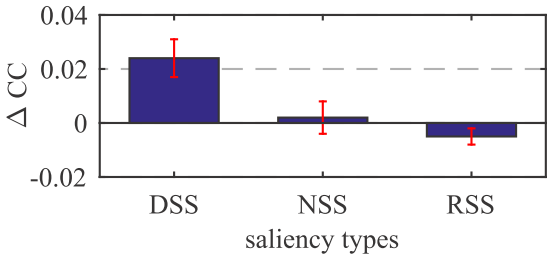


Fig. 6. Comparison of performance gain for VQMs weighted with DSS, NSS and RSS. The errorbars indicate 95% confidence level.

TABLE V

COMPARISON OF PERFORMANCE FOR VQMS WITH AND WITHOUT DSS WEIGHTING. (A) LINEAR CORRELATION COEFFICIENT (CC); AND (B) SPEARMAN RANK ORDER CORRELATION COEFFICIENT (SROCC)

	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.654	0.481	0.539	0.401	0.539
DSS-PSNR	0.688	0.457	0.559	0.455	0.567
SSIM	0.471	0.536	0.610	0.574	0.500
DSS-SSIM	0.567	0.534	0.642	0.603	0.544
ViS3	0.855	0.822	0.789	0.751	0.826
DSS-ViS3	0.859	0.808	0.829	0.765	0.839
STMAD	0.802	0.796	0.908	0.820	0.823
DSS-STMAD	0.816	0.798	0.911	0.835	0.834
SMOVIE	0.790	0.734	0.730	0.634	0.740
DSS-SMOVIE	0.843	0.710	0.765	0.678	0.771
TMOVIE	0.842	0.745	0.796	0.818	0.823
DSS-TMOVIE	0.809	0.720	0.825	0.848	0.836
MOVIE	0.836	0.759	0.787	0.725	0.795
DSS-MOVIE	0.862	0.744	0.816	0.759	0.829
VSSIM	0.591	0.552	0.578	0.589	0.584
DSS-VSSIM	0.618	0.554	0.597	0.616	0.605

(a)

	Wireless	IP	H.264	MPEG-2	All Data
PSNR	0.621	0.472	0.473	0.383	0.523
DSS-PSNR	0.640	0.477	0.499	0.441	0.543
SSIM	0.522	0.470	0.656	0.561	0.525
DSS-SSIM	0.573	0.442	0.698	0.610	0.567
ViS3	0.839	0.792	0.769	0.736	0.816
DSS-ViS3	0.843	0.749	0.842	0.742	0.833
STMAD	0.809	0.776	0.902	0.846	0.825
DSS-STMAD	0.819	0.776	0.898	0.857	0.830
SMOVIE	0.793	0.705	0.707	0.691	0.727
DSS-SMOVIE	0.800	0.692	0.693	0.734	0.755
TMOVIE	0.811	0.719	0.780	0.817	0.806
DSS-TMOVIE	0.775	0.667	0.810	0.852	0.808
MOVIE	0.811	0.716	0.766	0.773	0.789
DSS-MOVIE	0.830	0.701	0.789	0.816	0.809
VSSIM	0.582	0.548	0.572	0.577	0.587
DSS-VSSIM	0.613	0.555	0.581	0.603	0.606

(b)

both with $P<0.05$ at 95% confidence level. Based on the observed trend, we may conclude that modelling saliency in VQMs should target DSS rather than NSS. The inadequate performance gain obtained with NSS is possibly caused by two reasons. One reason is that some VQMs are already good at capturing NSS and, as a consequence, do not benefit from the addition of NSS (i.e., saturation effect in saliency-based optimization). The other reason might be that in some demanding conditions, NSS map does not sufficiently reflect the distraction power of artifacts occurring in some non-salient region, and therefore weighting a VQM with NSS might unhelpfully downplay the importance of distortion in this region.

V. THE INTEGRATION OF DSS IN VQMS: STATISTICS AND OPTIMIZATION

Granted that DSS rather than NSS is beneficial for VQMs, we further evaluate to what extent the actual amount of performance gain (when adding DSS to VQMs) changes for different VQMs and distortion types. Knowing the trends of such variation (i.e., building a benchmark) in performance gain is of high practical relevance to the application of saliency in VQMs.

A. Performance Gain and Statistical Significance

Table V shows the performance (in terms of CC and SROCC) of individual VQMs (with and without DSS weighting) when accessing different types of video distortion. Each entry in the table represents the performance of a VQM (with or without saliency weighting) on a subset of the LIVE database (i.e., a total of 40 data points for Wireless, 30 data points for IP, 40 data points for H.264 and 40 data points for MPEG-2). In general, this table demonstrates that there is indeed a gain in performance when adding DSS in VQMs. For the vast majority of cases, the performance of a DSS weighted VQM is higher than its original metric. However, the actual amount of such improvement varies, e.g., the performance gain of DSS-SSIM over SSIM for Wireless is 0.096 in terms of CC; whereas the difference in CC between DSS-TMOVIE and TMOVIE is -0.033 (but not necessarily meaningless). To verify whether the numerical difference in performance between a DSS weighted VQM and the same VQM without DSS is statistically significant, hypothesis testing is performed. As suggested in [53], the test is based on the residuals between DMOS and the outputs of a VQM (hereafter, referred to as M-DMOS residuals). Before being able to run an appropriate statistical significance test, we evaluate the assumption of normality of the M-DMOS residuals. The results of the test for normality are summarised in Table VI. As in all cases, paired M-DMOS residuals (i.e., two sets of residuals are compared: one is from the original VQM and one is from its DSS weighted version) are both normal, a paired samples t-test is performed (as used in [55]). The t-test results are given in Table VII, and show that in most cases the difference in performance between a VQM and its DSS weighted version is

TABLE VI

NORMALITY OF M-DMOS RESIDUALS. "1" REPRESENTS THE NORMAL DISTRIBUTION AND "0" REPRESENTS THE NON-NORMAL DISTRIBUTION

	Wireless	IP	H.264	MPEG-2
PSNR	1	1	1	1
DSS-PSNR	1	1	1	1
SSIM	1	1	1	1
DSS-SSIM	1	1	1	1
ViS3	1	1	1	1
DSS-ViS3	1	1	1	1
STMAD	1	1	1	1
DSS-STMAD	1	1	1	1
SMOVIE	1	1	1	1
DSS-SMOVIE	1	1	1	1
TMOVIE	1	1	1	1
DSS-TMOVIE	1	1	1	1
MOVIE	1	1	1	1
DSS-MOVIE	1	1	1	1
VSSIM	1	1	1	1
DSS-VSSIM	1	1	1	1

(a) (b)

Fig. 7. Illustration of the appearance of artifacts at low quality IP distortion. (a) A sample frame distorted with IP. (b) Saliency map (i.e., DSS) of (a).

TABLE VII

T-TEST RESULTS OF THE M-DMOS RESIDUALS. "1" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT. "0" MEANS THAT THE DIFFERENCE IS NOT SIGNIFICANT

	Wireless	IP	H.264	MPEG-2
PSNR vs. DSS-PSNR	1	1	1	1
SSIM vs. DSS-SSIM	1	1	0	1
ViS3 vs. DSS-ViS3	1	1	1	1
STMAD vs. DSS-STMAD	1	1	1	1
SMOVIE vs. DSS-SMOVIE	1	1	1	1
TMOVIE vs. DSS-TMOVIE	1	1	1	1
MOVIE vs. DSS-MOVIE	1	1	1	1
VSSIM vs. DSS-VSSIM	1	1	1	1

statistically significant. This implies that the addition of DSS in VQMs consistently makes a meaningful impact on their reliability of performance.

B. Variation in Performance Gain

To further comprehend the impact of the type of distortion and the kind of VQM on the changes of the performance gain achieved by adding DSS to VQMs, we re-arrange the entries of Table V with a focus on the increase in performance of an VQM when assessing a given distortion type. Table VIII illustrates the performance gain (expressed by ΔCC) of a DSS weighted VQM over its original metric when accessing Wireless, IP, H.264 and MPEG-2. It shows that MPEG-2 ($<\Delta CC>= 0.030$) benefit most from adding DSS for quality prediction, followed by Wireless ($<\Delta CC>= 0.027$) and H.264 ($<\Delta CC>= 0.026$); whereas there is a negetive effect for adding DSS to VQMs for assessing the quality of IP ($<\Delta CC>= -0.012$). In the case of a stimulus distorted with IP, especially at low quality, severe artifacts are often spread out over a large area of the scene as illustrated in Fig. 7. In such a scenario, the DSS map measured by eye-tracking, as shown in Fig. 7(b), may not be able to fully capture the perceptible artifacts and their impact on the judgement of video quality. As such, weighting a VQM with DSS may downplay the significance of potential distortion. This may explain the overall negative performance gain for IP, and tends to suggest that an optimised integration of saliency in VQMs may need to take this phenomenon into account.

TABLE VIII

PERFORMANCE GAIN (EXPRESSED BY THE INCREASE IN CC, I.E., ΔCC) BETWEEN A VQM AND ITS DSS WEIGHTED VERSION WHEN ASSESSING DIFFERENT DISTORTION TYPES (I.E., DENOTED AS ΔVQM)

	Wireless	IP	H.264	MPEG-2	mean
ΔPSNR	0.034	-0.024	0.020	0.054	0.021
ΔSSIM	0.096	-0.002	0.032	0.029	0.039
ΔViS3	0.004	-0.014	0.040	0.014	0.011
ΔSTMAD	0.014	0.002	0.003	0.015	0.009
ΔSMOVIE	0.053	-0.024	0.035	0.044	0.027
ΔTMOVIE	-0.033	-0.025	0.029	0.030	**0.000**
ΔMOVIE	0.026	-0.015	0.029	0.034	0.019
ΔVSSIM	0.027	0.002	0.019	0.027	0.019
mean	0.027	**-0.012**	0.026	0.030	

Table VIII also shows the performance gain (expressed by ΔCC) of a DSS weighted VQM over its original metric for individual VQM cases averaged over all distortion types. It shows that adding DSS results in a promising gain for all VQMs except for the case of TMOVIE. It is worth noting that for VQMs that already achieve a high prediction performance, such as ViS3 and STMAD, adding DSS still produces a significant increase in their performance (i.e., $<\Delta CC>= 0.011$ for ViS3 and $<\Delta CC>= 0.009$ for STMAD). In terms of the mean over all distortion types, TMOIVE does not benefit from the addition of DSS (i.e., $\Delta CC = 0$). This may be attributed to the fact that TMOVIE already contains sufficient saliency aspects in its metric design, e.g., it incorporates the estimate of motion, which is considered as a relevant cue in video saliency. Adding (possibly duplicated) saliency may be counter-productive in some cases as it may confuse the workings of the original VQM with built-in saliency.

C. Optimization: Proposed Integration Strategy

Section IV has demonstrated the superiority of DSS over NSS in improving VQMs. DSS, to some extent, reflects the interactions between natural scene and distortion, and therefore is observed to be more effective when adding a saliency term to VQMs. However, it is known that the recorded fixations

TABLE IX

COMPARISON OF PERFORMANCE (CC) FOR DIFFERENT VQMS AND THEIR CORRESPONDING SALIENCY-BASED VERSIONS, USING SIMPLE APPROACH AND PROPOSED APPROACH. THE LAST COLUMN PRESENTS THE PERFORMANCE GAIN (I.E., THE INCREASE IN CC (ΔCC)) OF A SALIENCY-BASED VQM OVER ITS ORIGINAL VERSION AVERAGED OVER ALL VQMS

	PSNR	SSIM	ViS3	STMAD	SMOVIE	TMOVIE	MOVIE	VSSIM	Averaged ΔCC
Original	0.539	0.500	0.826	0.823	0.740	0.823	0.795	0.584	-
with DSS + simple approach	0.567	0.544	0.839	0.834	0.771	0.836	0.829	0.605	0.024
with DSS + proposed approach	0.630	0.700	0.850	0.848	0.809	0.837	0.836	0.677	0.070

may not fully represent the entire human attentional behaviour [12], [17]. For example, the so called covert attention mechanisms, which refer to that of mentally focusing onto one of several possible sensory stimuli (without necessarily moving the eyes) [12], [17], may not be included in the DSS. This means that DSS may not be able to completely capture the attentional power of perceptible distortion, i.e., some artifacts in the visual field may be perceived but covertly attended (without any recorded fixations in the DSS). To address this phenomenon, we propose a more sophisticated integration strategy that better takes into account the attentional power of distortion. In [57], this idea has been initially explored for improving image quality prediction. We now extend it to a spatiotemporal framework for video quality assessment.

1) Proposed Approach: We now consider how to use the above concept to improve the formula expressed in (2). For each per-frame distortion map (PFDM) computed by a VQM, instead of using SM as a weighting factor, we now use two components: the captured DSS (i.e., denoted as α) and the uncaptured attentional power of distortion (i.e., denoted as β) to produce a local weighting factor ω. Given a pixel location (i, j), ω is defined as:

$$\omega_S(i, j) = f(\alpha, \beta_S) \tag{3}$$

$$\omega_T(i, j) = f(\alpha, \beta_T) \tag{4}$$

where ω_S (or ω_T) denotes the weighting factor for spatial PFDM (or temporal) portion of the PFDM, β_S (or β_T) denotes the uncaptured attentional power of spatial (or temporal) distortion. In this paper, β is modelled using an information theory based approach. This approach treats HVS as an optimal information extractor [58]; and β is considered to be proportional to the perceived information of distortion.

Based on the principle in [59], the perceived information I of a stimulus can be modelled as the number of bits transmitted from this stimulus (with the stimulus power S) through the visual channel of the HVS (with the noise power C); and can be computed as:

$$I = \frac{1}{2}\log(1 + \frac{S}{C}) \tag{5}$$

If we simply consider the distortion as the input stimulus, the perceived information of distortion can now be measured by the above formula. In such a scenario, the component S/C is analogous to the power of the locally measured distortion using PFDM. Due to the fact that HVS is not sensitive to pixel-level variations [60], the implementation of the algorithm is thus performed on the basis of a local patch of 45×45 pixels (about 2° visual angle in our experiment). Thus, (5) can be further defined as:

$$I_{S,P} = \frac{1}{2}\log(1 + \sigma_{s,p}^2) \tag{6}$$

$$I_{T,P} = \frac{1}{2}\log(1 + \sigma_{t,p}^2) \tag{7}$$

where $\sigma_{s,p}^2$ (or $\sigma_{t,p}^2$) estimates the power of the local spatial (or temporal) distortion within the patch P centred at a given pixel (i, j) in the PFDM; and $\sigma_{s,p}$ (or $\sigma_{t,p}$) denotes the standard deviation of P.

Moreover, our algorithm is motivated by the significant findings in [61] that each perceptible artifact suppresses each other artifact's effect especially for those with close proximity. This so-called surround suppression effect (SSE) is used to approximate the proportional relationship between β and I, where the effect of I is suppressed by its local neighbourhood. Thus, β can be defined as:

$$\beta_{S,P} = \frac{I_{S,P}}{\bar{I}_S} \tag{8}$$

$$\beta_{T,P} = \frac{I_{T,P}}{\bar{I}_T} \tag{9}$$

where $\bar{I}_S$ (or $\bar{I}_T$) represents the averaged spatial (or temporal) attentional power of distortion surrounding the local patch P. In this paper, the vicinity is defined as the Moore neighbourhood of the local patch P (i.e., the set of eight patches P_k ($k = 1$ to 8) of the same size which share a vertex or edge with P).

Finally, we combine α and β using a simple multiplication operator, resulting in the spatial and temporal weighting factors:

$$\omega_S(i, j) = \alpha^m \cdot {\beta_S}^n \tag{10}$$

$$\omega_T(i, j) = \alpha^x \cdot {\beta_T}^y \tag{11}$$

where $m > 0$, $n > 0$, $x > 0$ and $y > 0$ are parameters to adjust the relative importance of different components. We set $m = n = x = y = 1$ in our experiment for simplification. Tuning the parameters may improve the algorithm; however it goes beyond the merits of this paper. Once ω is achieved, we use it to replace the term SM in (2). Noted that for VQMs that only perform in the spatial channels (e.g., PSNR and SSIM), only ω_S is calculated.

2) Validation of the Approach: For each VQM, we compare its DSS-weighted version using (2) (referred to as simple approach) and that using the proposed approach. Table IX shows the performance (i.e., in terms of CC, SROCC exhibits the same trend of changes as CC and thus is not included here) in each case. It shows that the proposed approach performs

TABLE X

COMPARISON OF PERFORMANCE FOR VQMS WEIGHTED WITH MEASURED SALIENCY (DSS) AND MODELLED SALIENCY (BASED ON FIVE SALIENCY MODELS, NAMELY SR, PQFT, GBVS, SDSR AND CA). TWO WEIGHTING APPROACHES ARE USED: SIMPLE APPROACH (METH-1) AND PROPOSED APPROACH (METH-2)

	original	DSS		GBVS		SDSR		PQFT		CA		SR	
		Meth-1	Meth-2	Meth-1	Meth-2	Meth-1	Meth-2	Meth-1	Meth-2	Meth-1	Meth-2	Meth-1	Meth-2
PSNR	0.539	0.567	0.630	0.510	0.578	0.540	0.579	0.566	0.607	0.501	0.574	0.516	0.579
SSIM	0.500	0.544	0.700	0.575	0.689	0.641	0.661	0.582	0.717	0.538	0.690	0.580	0.694
ViS3	0.826	0.839	0.850	0.828	0.843	0.833	0.846	0.837	0.848	0.822	0.835	0.836	0.845
STMAD	0.823	0.834	0.848	0.824	0.835	0.825	0.828	0.838	0.850	0.817	0.821	0.830	0.843
SMOVIE	0.740	0.771	0.809	0.759	0.771	0.741	0.805	0.776	0.796	0.754	0.787	0.737	0.777
TMOVIE	0.823	0.836	0.837	0.819	0.822	0.823	0.817	0.786	0.797	0.810	0.816	0.812	0.813
MOVIE	0.795	0.829	0.836	0.816	0.817	0.796	0.813	0.798	0.802	0.799	0.803	0.791	0.801
VSSIM	0.584	0.605	0.686	0.598	0.666	0.579	0.671	0.602	0.679	0.574	0.683	0.595	0.687

consistently better in each comparison. A paired samples t-test analysis (preceded by a test for the assumption of normality) was further performed, selecting the integration approach as the independent variable and the performance as the dependent variable. The result shows that the proposed approach statistically significantly outperforms the simple approach with $P<0.05$ at 95% confidence level.

VI. EVALUATION OF MODELLED SALIENCY IN VQMS

We evaluate whether a saliency model, at least with the current soundness of visual saliency modelling, can sufficiently benefit VQMs in comparison to the gain yielded by DSS. Our evaluation is carried out with five saliency models, namely SR [62], PQFT [63], GBVS [64], SDSR [65] and CA [66]. Each saliency model is integrated in VQMs using both the simple and proposed weighting methods.

These saliency models have been extensively studied in the context of image quality assessment, and demonstrated to be among the best performing saliency models in terms of producing consistent performance gain for image quality prediction [52]. It is, therefore, worth investigating the added value of these saliency models in VQMs. Note that some models, such as SR and CA are specifically designed for still images; and models, such as GBVS, PQFT and SDSR account for both spatial and temporal aspects in saliency modelling.

Table X shows the comparison of the performance (i.e., expressed in terms of CC, SROCC exhibits the same trend of changes as CC, and thus is not included here) of VQMs using both the simple and proposed approaches. Fig. 8 further illustrates the averaged performance gain (i.e., expressed in terms of ΔCC) for both approaches. The performance that can be achieved by adding DSS in VQMs is also included as a reference. In terms of using the simple method, the table tends to indicate that there does not exist a saliency model that can consistently benefit all VQMs. The addition of a saliency model may improve the performance of a specific VQM, while running the risk of deteriorating other VQMs' performance. For example, adding SR to SSIM can boost its performance by 0.08 in CC, but this saliency model significantly degrades the performance of PSNR. Existing saliency models, on average, hardly improve the prediction performance of VQMs (i.e., ΔCC is about 0.01), compared to the benefit of DSS (i.e., ΔCC is around 0.025). Instead, when using the proposed

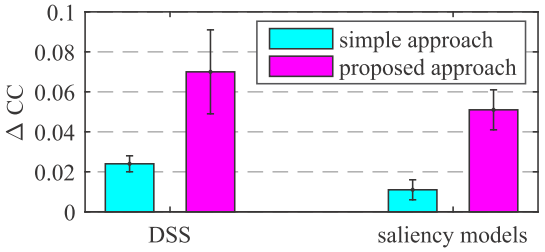


Fig. 8. Performance gain obtained by adding DSS versus saliency models to VQMs using both simple and proposed integration approaches. The errorbars indicate 95% confidence interval.

approach, firstly, saliency models can benefit VQMs in a more consistent way. For example, almost all saliency-based VQMs outperform their original metics. Secondly, on average, as shown in Fig. 8 the performance again obtained by adding saliency models (i.e., ΔCC is 0.051) is significantly increased, but is still lower than the gain achieved by adding DSS. Fig. 8 suggests that compared to the gain of "ground truth" saliency, modelling saliency in VQMs contains sufficient headroom for further improvement.

VII. DISCUSSION

Our statistical evaluation provides general insight into the benefits of saliency in VQMs. The approach (either simple or proposed) used for combining saliency and an VQM is universally applicable. This means that this combination method can be applied to all kinds of VQMs that measure distortions locally. One should realise though that tuning a specific VQM with a specifically designed saliency weighting, e.g., using saliency aspects to optimise the contrast sensitivity function (CSF) contained in a VQM [35], may produce superior improvement for that particular VQM. There is, however, no guarantee that this specifically designed approach can be easily implemented in other VQMs or improve their performance. In terms of conducting a rigorous comparative study, a generic saliency integration method is highly required.

Our empirical evidence shows that NSS is not suitable for VQMs, and it is DSS that should be included in VQMs. This conclusion was drawn by using the simple saliency integration approach. One may wonder whether this conclusion holds when the proposed integration approach is used. We repeated the experiment as described in Section IV, using the proposed

integration approach. Our experimental results show that the proposed approach increases the benefits of NSS in absolute terms, while maintaining the relative difference in the gain between NSS and DSS. In addition, from a practical point of view, it may be unrealistic to calculate saliency from the reference video simply because the reference is not always available in many real-world applications.

VIII. Conclusion

In this paper, we investigated saliency and its use in objective video quality assessment. To obtain reliable "ground truth" saliency for video quality research, we proposed a refined experimental methodology and conducted a large-scale eye-tracking experiment. In our experiment, a large number of human observers freely looked to a diverse range of video stimuli distorted with different types of distortion at various levels of degradation. We applied dedicated control mechanisms with the aim to overcome bias that potentially exists in related studies.

Based on the "ground truth" data of saliency, we performed an exhaustive statistical evaluation to assess the effects of saliency on the reliability of VQMs. We found a tendency that adding DSS rather than NSS to a VQM improved its performance in predicting perceived video quality. Based on this evidence, the added value of DSS in VQMs was further assessed. This evaluation shows that there is a statistically significant gain in the performance for all VQMs when adding DSS. The extent of the performance gain, however, tends to depend on the specific distortion type assessed and the VQM under test. We also investigated integrating saliency in VQMs in a perceptually more relevant way, and devised a generic approach that can optimise the use of saliency in VQMs.

Finally, we applied several state-of-the-art computational models of visual saliency in VQMs, and assessed their capabilities in improving the VQM's performance. Quantitative results tend to show that compared to the improvement achieved by using eye-tracking data, modelling saliency in the context of video quality requires further investigation.

References

[1] S. Winkler, "Video quality measurement standards: Current status and trends," in *Proc. 7th Intl. Conf. ICICS*, Macau, China, 2009, pp. 848–852.

[2] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[3] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *Proc. Handbook Video Databases: Design Appl.*, 2003, pp. 1041–1078.

[4] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 253–265, Apr. 2009.

[5] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 266–279, Apr. 2009.

[6] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[7] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[8] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. 18th IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 2505–2508.

[9] P. V. Vu and D. M. Chandler, "ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, p. 013016, Feb. 2014.

[10] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 121–132, 2004.

[11] M. I. Posner and S. E. Petersen, "The attention system of the human brain," *Annu. Rev. Neurosci.*, vol. 13, pp. 25–42, 1990.

[12] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2013.

[13] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *Science*, vol. 315, no. 5820, pp. 1860–1862, 2007.

[14] E. Niebur and C. Koch, "Computational architectures for attention," in *Proc. Attentive Brain*, 1998, pp. 163–186.

[15] J. H. Fecteau and D. P. Munoz, "Salience, relevance, and firing: A priority map for target selection," *Trends Cognit. Sci.*, vol. 10, no. 8, pp. 382–390, 2006.

[16] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.

[17] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[18] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[19] W. Y. L. Akamine and M. C. Q. Farias, "Video quality assessment using visual attention computational models," *J. Electron. Imag.*, vol. 23, no. 6, p. 061107, 2014.

[20] B. Fu, Z. Lu, X. Wen, L. Wang, and H. Shao, "Visual attention modeling for video quality assessment with structural similarity," in *Proc. 16th Int. Symp. Wireless Pers. Multimedia Commun.*, Jun. 2013, pp. 1–5.

[21] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency inspired full-reference quality metrics for packet-loss-impaired video," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 81–88, Mar. 2011.

[22] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj, "Salient motion features for video quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 948–958, Apr. 2011.

[23] J. You, J. Korhonen, and A. Perkis, "Attention modeling for video quality assessment: Balancing global quality and local quality," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010, pp. 914–919.

[24] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency based objective quality assessment of decoded video affected by packet losses," in *Proc. 15th IEEE Int. Conf. Image Process.*, San Diego, CA, USA, Oct. 2008, pp. 2560–2563.

[25] M. C. Q. Farias and W. Y. L. Akamine, "On performance of image quality metrics enhanced with visual attention computational models," *Electron. Lett.*, vol. 48, no. 11, pp. 631–633, May 2012.

[26] Y. Li, X. Guo, and H. Wang, "Spatio-temporal quality pooling adaptive to distortion distribution and visual attention," in *Proc. Int. Conf. Vis. Commun. Image Process.*, Dec. 2015, pp. 1–4.

[27] W. Zhao, L. Ye, J. Wang, and Q. Zhang, "No-reference objective stereo video quality assessment based on visual attention and edge difference," in *Proc. 2015 IEEE Adv. Inf. Technol. Electron. Autom. Control Conf.*, Dec. 2015, pp. 523–526.

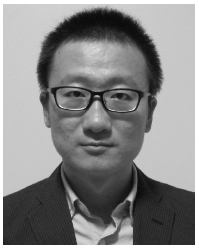
[28] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[30] H. Alers, J. A. Redi, and I. Heynderickx, "Quantifying the importance of preserving video quality in visually important regions at the expense of background content," *Signal Process., Image Commun.*, vol. 32, pp. 69–80, Mar. 2015.

[31] U. Engelke, R. Pepion, P. Le Callet, and H.-J. Zepernick, "Linking distortion perception and visual saliency in H.264/AVC coded video containing packet loss," *Proc. SPIE, Vis. Commun. Image Process.*, vol. 7744, p. 774406, Jul. 2010.

[32] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric," *Signal Process., Image Commun.*, vol. 25, no. 7, pp. 547–558, Aug. 2010.
[33] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Do video coding impairments disturb the visual attention deployment?" *Signal Process., Image Commun.*, vol. 25, no. 8, pp. 597–609, 2010.
[34] M. Pinson, M. Sullivan, and A. Catellier, "A new method for immersive audiovisual subjective testing," in *Proc. 8th Int. Workshop Video Process. Quality Metrics Consumer Electron.*, 2014, pp. 1–6.
[35] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 200–213, Jan. 2014.
[36] W. Zhang and H. Liu. *An Eye-tracking Database for Video Quality Assessment*, accessed on 2017. [Online]. Available: https://sites.google.com/site/vaqatoolbox/
[37] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[38] J. Krauth, *Experimental Design: A Handbook and Dictionary for Medical and Behavioral Research*, vol. 14. New York, NY, USA: Elsevier, 2000.
[39] A. G. Greenwald, "Within-subjects designs: To use or not to use?" *Psychol. Bull.*, vol. 83, no. 2, p. 314, 1976.
[40] G. Keren, "Between-or within-subjects design: A methodological dilemma," in *A Handbook For data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, NJ, USA: Erlbaum, 1993, p. 257.
[41] *Methodology for the Subjective Assessment of the Quality of Television Pictures, Recommendation*, document ITU-R BT-500.13, 2012.
[42] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," Comput. Sci. Artif. Intell. Lab, Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, 2012.
[43] A. Borji and L. Itti. (2015). "CAT2000: A large scale fixation dataset for boosting saliency research." [Online]. Available: https://arxiv.org/abs/1505.03581
[44] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye Tracking Res. Appl.*, Florida, FL, USA, 2000, pp. 71–78.
[45] D. Cousineau and S. Chartier, "Outliers detection and treatment: A review," *Int. J. Psychol. Res.*, vol. 3, no. 1, pp. 58–67, 2010.
[46] V. Barnett and T. Lewis, *Outliers in Statistical Data* (Wiley Series in Probability and Statistics), vol. 1. 2nd ed., Chichester, U.K.: Wiley, 1984, 1984.
[47] U. Engelke *et al.*, "Comparative study of fixation density maps," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1121–1133, Mar. 2013.
[48] S. Winkler and S. Ramanathan, "Overview of eye tracking datasets," in *Proc. 5th Int. Workshop QoMEX*, Klagenfurt, Austria, 2013, pp. 212–217.
[49] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychol. Rev.*, vol. 113, no. 4, p. 766, 2006.
[50] T. Judd, F. Durand, and A. Torralba, "Fixations on low-resolution images," *J. Vis.*, vol. 11, no. 4, p. 14, 2011.
[51] W. Zhang, A. Borji, F. Yang, P. Jiang, and H. Liu, "Studying the added value of computational saliency in objective image quality assessment," in *Proc. IEEE Vis. Commun. Image Process. Conf.*, Dec. 2014, pp. 21–24.
[52] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, Jun. 2016.
[53] Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective models of video quality assessment," VQEG, Tech. Rep., 2000.
[54] S. Winkler, "Vision models and quality metrics for image processing applications," Ph.D. dissertation, Dept. Elect. Eng., Univ. Lausanne, Lausanne, Switzerland, 2000.
[55] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 971–982, Jul. 2011.
[56] D. C. Montgomery, *Applied Statistics and Probability for Engineers*, 6th ed. Hoboken, NJ, USA: Wiley, 2013.
[57] W. Zhang, J. Talens-Noguera, and H. Liu, "The quest for the integration of visual saliency models in objective image quality assessment: A distraction power compensated combination strategy," in *Proc. 22nd IEEE Int. Conf. Image Process.*, Quebec City, QC, Canada, Sep. 2015, pp. 1250–1254.
[58] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.
[59] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. IEEE Intl. Conf. Image Process.*, Oct. 2006, pp. 2945–2948.
[60] Z. Wang and A. C. Bovik, "Modern image quality assessment," in *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2. San Rafael, CA, USA: Morgan & Claypool, Jan. 2006, no. 1, pp. 1–156.
[61] L. K. Chan and W. G. Hayward, "Dimension-specific signal modulation in visual search: Evidence from inter-stimulus surround suppression," *J. Vis.*, vol. 12, no. 4, p. 10, 2012.
[62] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
[63] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.
[64] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. 20th Conf. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 545–552.
[65] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, 2009.
[66] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.



Wei Zhang (S'14) received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Informatics, Cardiff University. His research interests include visual quality assessment, video processing, and human visual perception.

Hantao Liu (S'07–M'11) received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2011. He is currently an Assistant Professor with the School of Computer Science and informatics, Cardiff University, Cardiff, U.K. His research interests include visual media quality assessment, visual attention modeling and applications, visual scene understanding, and medical image perception. He is currently serving for the IEEE MMTC, as the Chair of the Interest Group on Quality of Experience for Multimedia Communications. He is an Associate Editor of the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS.